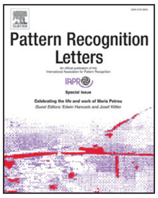
Pattern Recognition Letters 73 (2016) 19–25

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec



Cascaded regression with sparsified feature covariance matrix for facial landmark detection[☆]

Enrique Sánchez-Lozano, Brais Martinez, Michel F. Valstar[*]

School of Computer Science, University of Nottingham, Nottingham, NG8 1BB, UK

ARTICLE INFO

Article history:
Received 8 April 2015
Available online 13 January 2016

Keywords:
Supervised descent method
Cascaded regression
Facial point localisation

ABSTRACT

This paper explores the use of context on regression-based methods for facial landmarking. Regression based methods have revolutionised facial landmarking solutions. In particular those that implicitly infer the whole shape of a structured object have quickly become the state-of-the-art. The most notable exemplar is the Supervised Descent Method (SDM). Its main characteristics are the use of the cascaded regression approach, the use of the full appearance as the inference input, and the aforementioned aim to directly predict the full shape. In this article we argue that the key aspects responsible for the success of SDM are the use of cascaded regression and the avoidance of the constrained optimisation problem that characterised most of the previous approaches. We show that, surprisingly, it is possible to achieve comparable or superior performance using only landmark-specific predictors, which are linearly combined. We reason that augmenting the input with too much context (of which using the full appearance is the extreme case) can be harmful. In fact, we experimentally found that there is a relation between the data variance and the benefits of adding context to the input. We finally devise a simple greedy procedure that makes use of this fact to obtain superior performance to the SDM, while maintaining the simplicity of the algorithm. We show extensive results both for intermediate stages devised to prove the main aspects of the argumentative line, and to validate the overall performance of two models constructed based on these considerations.

1. Introduction

Structured object detection is an active research area in Computer Vision, where the aim is to describe the shape of an object by locating its parts. Facial landmark detection is a prime example of this, and it is a key step in many applications such as face recognition or facial expression recognition, where the alignment step based on the location of the parts is crucial to achieve a good performance.

Existing facial landmark detection approaches are commonly divided into part-based and holistic approaches. Holistic approaches are mostly restricted to the Active Appearance Models family ([5,11]). They represent the full face appearance, and are typically generative. Facial landmarking results in this case as a by-product of the dense reconstruction of the face appearance. Instead, part-based models are characterised by representing the face as a constellation of patches, each centred around the facial landmarks. They are typically discriminative [15], although it is also possible to use part-based generative models [17]. While generative methods are capable of attaining very precise results when the search is initialised close to the solution [16], discriminative methods provide better robustness. In this article we focus on part-based discriminative models, as they are the most widely used.

Many of the existing works on part-based facial landmarking can be cast in the Constrained Local Models (CLM) framework[1] introduced by [15]. The CLM framework devises landmark detection as the iterative alternation between two steps, response map construction and response maximisation. Response maps encode the likelihood of any given image location of being the true landmark location, and a different response map is constructed for each landmark. Many works used classifiers to create these

[☆] This paper has been recommended for acceptance by Dr. S. Dutta Roy.

[*] Corresponding author. Tel.: +44 115 823 2287; fax: +44 115 951 4254.
E-mail address: Michel.Valstar@nottingham.ac.uk (M.F. Valstar).
URL: http://www.cs.nott.ac.uk/~mfv (M.F. Valstar)

[1] The term Constrained Local Model was previously introduced by [7] prior to the work by [15]. Furthermore, it has become somewhat common to refer to the specific approach proposed in [15] as the CLM, while their method was introduced only as a particular instance of the CLM framework. In this article we refer to CLM as the general framework rather than to any specific methodology.

http://dx.doi.org/10.1016/j.patrec.2015.11.014

landmarks (e.g. [1–3,15]). A probabilistic classifier (e.g., a logistic regressor) can be trained to distinguish the true landmark location from surrounding locations. At test time, the classifier can be evaluated over a region of interest in a sliding window manner. The response map is then constructed using the predicted likelihoods. The response maximisation step consists of finding the valid shape maximising the combined per-landmark responses. Thus, this step is a maximisation constrained by the shape model.

The shape fitting step is very challenging, and it contains multiple local minima. Thus, many authors have focused their efforts on improving this step. For example, [15] attained real-time reliable fitting by using a Mean Shift-constrained optimisation. However, the Mean Shift optimisation is prone to converge at local maxima, especially for the flexible shape parameters, responsible for expressions. To overcome this, [3] proposed a variation of RANSAC, so that a very large number of solutions were generated using training set exemplars. The highest-scoring exemplars were linearly combined into the final solution. [1] instead used discriminatively trained regressors to find adequate increments to the shape parameters, and [2] proceeded by training a generative model of the response maps and then using it to perform the maximisation.

Recent years have seen the appearance of works employing regressors instead of classifiers to exploit local appearance [18]. It was soon shown that the regressors resulted in improved response maps and hence better global performance (e.g. [6,10]). However, a constrained optimisation problem was still necessary in order to enforce shape consistency, consequently hindering performance. Further performance improvement was attained by considering regressors trained to directly infer the full shape increments necessary to move from the current shape estimate to the ground truth. That is to say, instead of using the appearance of a single landmark to predict only the location of this landmark, the full appearance is used to predict the entire shape, eliminating the need for a subsequent step enforcing shape consistency. This was pioneered by [4], who also proposed the use of cascaded regression [8] to this end. However, it was the Supervised Descent Method (SDM) [19] that became the de-facto state of the art. While they maintained the main concepts of [4], they simplified the method by using Least Squares for regression, and concatenated per-landmark HOG features as their feature representation. This resulted in a very simple algorithm capable of attaining the best performance to date (only 4 matrix multiplications are involved, not counting feature extraction!).

Is thus an important line of investigation to analyse what the key advantages are of the SDM with respect to other methods. Several factors characterise the algorithm: the cascaded regression, the implicit use of context (i.e., the concatenation of all the local descriptors into a single feature vector), and the direct prediction of the shape. Each can be argued to have merit. The cascaded regression allows for combined robustness and precision, the use of context provides an input with augmented descriptive power, and the direct shape increment prediction removes the need for subsequent complex optimisation steps.

We argue that using only two of these components, to wit the cascaded regression and the direct estimation of the shape, is sufficient to produce similar or even better results to those of the SDM. That is to say, if these two aspects are respected, similar performance can be attained with and without context. We further investigate to which extent the use of context within the input features is necessary, exploring intermediate solutions between landmark-independent predictions and the SDM approach. In order to eliminate context from the regression models, we resort to the sparsification of the feature covariance matrix. We show experiments highlighting the relation between the amount of context used (i.e., the sparseness of the feature covariance matrix), and the variability of the data in terms of factors such as the head pose, image quality, facial expressions or identity. Finally, we use this relation to build a variant of the SDM algorithm with decreasingly sparse matrices at each iteration. This algorithm can be very easily implemented given an SDM implementation, has less computational complexity, and achieves superior performance in practise. We use the LFPW, Helen, AFW and IBUG datasets (see Section 6 for details) to validate the analysis and to show practical performance of the solution derived from it.

A previous version of this manuscript appeared in [14]. The work presented in this article differs from it in that we provide a more complete interpretation and mathematical derivation to justify the matrix sparsification, provide a link between the benefits of sparsification and data variance that was missing in the previous version, and we link the success of direct regression-based methods with the avoidance of constrained optimisation.

The contributions of this work can thus be summarised as:

- We analyse which are the key methodological aspects behind the performance success of the SDM.
- We show that, surprisingly, we achieve superior performance to the standard SDM when encoding no context within the input features.
- We show that there is an inverse correlation between the benefits of using context and the variance of the input data.
- Based on these observations, we devise a simple yet effective extension of the SDM, where each regressor uses an optimal amount of context within the input features. The resulting method is shown to outperform SDM.

2. Cascaded linear regression

Let $\mathbf{I}$ be a face image, for which we want to estimate the ground truth shape $\mathbf{s}^g$, consisting of n facial landmarks (thus being a $2n$-dimensional vector). Let $\mathbf{s}$ be an estimation of the location of these points, then $\boldsymbol{\phi}(\mathbf{I},\mathbf{s}) \in \mathbb{R}^{p\times 1}$, with p the dimension of the feature space, represents the features extracted around the positions defined by $\mathbf{s}$ within image $\mathbf{I}$. The feature vector is constructed by extracting a HOG descriptor at a small patch centred around each landmark, and then concatenating features of all patches into a single feature vector. The regression target is defined as $\delta = \mathbf{s}^g - \mathbf{s}$. That is to say, δ is the increment necessary to move from the current estimate $\mathbf{s}$ to the ground truth shape $\mathbf{s}^g$. It is then possible to define a linear regressor $\{\mathbf{R},\mathbf{b}\} \in \{\mathbb{R}^{2n\times p}, \mathbb{R}^{2n\times 1}\}$ tasked with translating image features into shape increments. Specifically, the increment δ is estimated as $\mathbf{R}\boldsymbol{\phi}(\mathbf{I},\mathbf{s}) + \mathbf{b}$ and the updated shape estimate is computed as $\delta + \mathbf{s}$. This linear regressor can be expressed in a more compact form by defining $\tilde{\boldsymbol{\phi}}(\mathbf{I},\mathbf{s})$ as the result of adding a one to the end of $\boldsymbol{\phi}(\mathbf{I},\mathbf{s})$. Then, $\tilde{\mathbf{R}}$ is defined as a $\mathbb{R}^{2n\times p+1}$ matrix, so that:

$$\mathbf{R}\boldsymbol{\phi}(\mathbf{I},\mathbf{s}) + \mathbf{b} = \tilde{\mathbf{R}}\tilde{\boldsymbol{\phi}}(\mathbf{I},\mathbf{s}) \tag{1}$$

The data variance is in practise too large to attain an accurate prediction of the true shape using only a single prediction made by one single regressor. In the SDM, this limitation is overcome through the use of the cascaded regression. The idea is to sequentially apply a set of regressors rather than using a single one. At test time, an initial shape estimate $\mathbf{s}^0$ is computed using the face detection bounding box. Then, the cascaded regression produces a sequence of estimates as $\mathbf{s}^k = \mathbf{s}^{k-1} + \tilde{\mathbf{R}}^k\tilde{\boldsymbol{\phi}}(\mathbf{I},\mathbf{s}^{k-1})$. If the cascade has N iterations, then $\mathbf{s}^N$ is the estimate of $\mathbf{s}^*$.

The training of the cascade starts with a *data augmentation* strategy [8], which proceeds by generating m different initial shapes for each of the n_{im} training images. These shapes can for example be generated by aligning a reference shape (e.g. the mean